Classification by Support Vector Machines

Santiago Velasco Forero
Advisor: Edgar Acuña
Mathematics Departament
University of Puerto Rico, Mayagüez Campus
Mayagüez, Puerto Rico
santi_vf@math.uprm.edu

March 10, 2003

Abstract

Support Vector Machines (SVMs) perform pattern recognition between two points classes by finding a decision surface determined by certain points the training set, termed Support Vectors (SV).

1 Introduction

Support Vector Machine (SVMs) have been introduced for Vapnik V.N. as a new technique for solving pattern recognition problem [CV95], [SV97], [FF97], [SS98], function estimation [VC71], times series analysis [KV97] and variance analysis [WW97]. The SVM algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in the sixties [VL63], [VC64]. What makes SVMs attractive is (a) the ability to condense the information contained in the training set, and (b) the use of families of decision surfaces of relatively low VC-dimension.

In the linear, separable case the key idea of a SVM can be explained in plain words. Given a training set S which contains points of either of two classes, a SVM separates the classes through a hyperplane determined by certain points of S, termed support vectors. In the separable case, this hyperplane maximizes the margin, or twice the minimum the distance of the either class from the hyperplane, and all support vectors lie at the same minimum distance from the hyperplane (and are thus termed support margin).

2 Theorical Overview

In this section, we recall the basic of the theory of SVM [CV95] in both the linear and nonlinear case.

2.1 Optimal separating hyperplane

In what follows we assume we are given a set S of points $x_i \in R^n$ with $i = 1, 2, ..., N$. Each point x_i belongs to either of two classes and thus is given a label $y_i \in -1, 1$. The goal is to establish of a hyperplane that divides S leaving all the point of the same class on the same side while maximizing the minimum distance between either of the two classes and the hyperplane. To this purpose we need some preliminary definitions.

Definition 1: The set S is *linearly separable* if there exist $w \in \mathbf{R}^n$ and $b \in \mathbf{R}$ such that:

$$\{\, w \cdot x_i + b \geq 1, \quad if \quad y_i = 1 w \cdot x_i + b \leq 1, \quad if \quad y_i = -1 \tag{1}$$

In compact notation, the two inequalities (1) can be rewritten:

$$y_i(w \cdot x_i + b) \geq 1, \quad i = 1, \ldots, l \tag{2}$$

The pair (w, b) defines a hyperplane of equations:

$$w \cdot x + b = 0 \tag{3}$$

named *separating hyperplane* .If we denote with $\|w\|$ the norm of w, the signed distance d_i of point x_i from the separating hyperplane (w, b) is given by:

$$d_i = \frac{w \cdot x_i + b}{\|w\|} \tag{4}$$

Combining inequality (2) and equation (3), for all $x_i \in S$ we have:

$$y_i d_i \geq 1/\|w\| \tag{5}$$

Therefore, $1/\|w\|$ is the lower bound on the distance between the points x_i and the separating hyperplane (w, b).

Definition 2: Given a separating hyperplane (w, b) for the linearly separable set S, the *canonical representation* of the separating hyperplane is obtained by rescaling the pair (w, b) into the pair (w', b') in such a way that the distance of the closest point equals $1/\|w'\|$.

In what follows we will assume that a separating hyperplane is always given the canonical representation and thus write (w, b) instead of (w', b').

Definition 3: Given a linearly separable set S, the optimal separating hyperplane (OSH) is the separating hyperplane which maximizes the distance of the closest point of S.

Since the distance of the closest point equals $1/\|w'\|$, the OSH can be regarded as the solution of the problem maximizing $1/\|w'\|$ subject of constraint (2), o:

$$\begin{array}{rc} \textit{Problem} & \mathbf{P1} \\ \textit{Minimize} & \frac{1}{2} w \cdot w, \\ \textit{Subject} & y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, ..., N \end{array}$$

If the pair (w, b) solve **P1**, then for at least one $x_i \in S$ we have $y_i(w \cdot x_i + b) = 1$. In particular, the implies that the solution of **P1** is always a separating hyperplane in the canonical representation. Moreover, the parameter b enters in the constraints but not in the function to be maximized

The quantity $2/\|w'\|$, which measure the distance between the two classes in the direction of w, is named *margen*. Hence, the OSH can also be seen as a separating hyperplane which maximizes the margin.

3 Linearly nonseparable case

If the set S is not linearly separable or one simply ignore whether or not the set S is linearly separable, the problem of searching for an OSH is meaningless. In this case, the previous analysis can be generalized by introducing N nonnegative variables $\xi = (\xi_1, \xi_2, ..., \xi_N)$ such that

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \tag{6}$$

for $i = 1, 2, ..., N$. If the point x_i satisfies inequality (2), the ξ_i are zero and (6) reduces to (2). Instead, if the point x_i does no satisfy inequality (2), the term $-\xi_i$ is added to the right hand side of (2) to obtain inequality (6). The generalized OSH is then rearded as the solution to:

$$\begin{array}{rc} \textit{Problem} & \mathbf{P2} \\ \textit{Minimize} & \frac{1}{2} w \cdot w + C \sum \xi_i \\ \textit{Subject} & y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \xi \geq 0 \end{array}$$

The term $C \sum \xi_i$, where the sum is for $i = 1, ...N$, can be thought of as some measure of the amount misclassification. Note that this term leads to a more robust solution , in the statistical sense, than the intuitively more appealing term $C \sum \xi_i^2$. In other words, the term $C \sum \xi_i$ makes the OSH less sensitive to the presence of outliers in the training set. The parameter C can be regarded as a regularization parameter. The OSH tends to maximize the minimum distance $1//\|w\|$ for small C, and minimized the number of misclassification points for large C. For intermediate values of C the solution of problem P2 trade errors for larger margin.

4 Nonlinear kernels

In most cases, linear separation in input space is a too restrictive hypothesis to be of practical use. Fortunately, the theory can be extended to nonlinear

separating surfaces by mapping the input points into feature points. If $x \in \mathbf{R}^n$ is a ponit, we let $\varphi(x)$ be the corresponding feature point with φ a mapping from $\mathbf{R}^n$ to certain space Z. We denote with φ_i the components of φ. Clearly, to an OSH in Z corresponds a nonlinear separating surface in input space. At first sight it might seem that this nonlinear surface cannot be determined unless the mapping φ is completely known. However, from the formulation of problem **P2** and the classification stage of equation (6), sit follows that φ enters only in the dot product between feature points in input, since

$$D_{ij} = y_i y_j \varphi(x_i) \cdot \varphi(x_j) \tag{7}$$

$$\bar{w} \cdot \varphi(x) + \bar{b} = \sum \bar{\alpha_i} y_i \varphi(x_i) \cdot \varphi(x) + \bar{b} \tag{8}$$

If we find an expression for the dot product in feature space which uses the points in input space only, that is:

$$\varphi(x_i) \cdot \varphi(x_j) = \mathcal{K}(x_i, x_j) \tag{9}$$

the full knowledge of φ is not necessary. The symmetric function $\mathcal{K}$ is called *kernel*. The nonlinear separating surface can be found as the solution of problem **P4** with $D_{ij} = y_i y_j \mathcal{K}(x_i, x_j)$, while the classification stage reduces to computing

$$sign(\sum \bar{\alpha_i} y_i \mathcal{K}(x_i, x_j) + \bar{b}) \tag{10}$$

Therefore, the extension of the theory to the nonlinear case is reduced to finding kernels with identify certain families of decision surfaces and can be written as in equation (9).

5 Experiments

We present the performances of the Support Vector Machine and the best single classifiers in 12 database:

Table 1: Result of the classification database

Database	SVMs	Best Classifiers
Iris	3.80	2(LDA)
Sonar	24.15	15.5(1-NN)
Glass	30.51	23.8(1-NN)
Heart-c	18.60	16.59(LDA)
Ionosphere	13.50	8.1(C4.5)
Crx	15.60	13.48(LDA)

References

[CV95] C. Cortes and V. Vapnik. Support vector network. *Machine Learning*, 20:1–25, 1995.

[FF97] E. Osuna, R. Freund and F.Girosi. Support vector machine: Training and aplications. *Technical Report Massachusetts Institute of Technology*, 1997.

[KV97] K.R. Muller, A. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen and V. Vapnik. Predicting time series with support vector machine. *Proceedings of the International Conference on Artifical Neural Networks*, page 999, 1997.

[SS98] A. Smola and B. Scholkolpf. A tutorial on support vector regression. *NeuroCOLT Technical Report Series*, 1998.

[SV97] H. Drucker, J.C. Burges, L. Kaufman, A. Smola and V. Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9:155–161, 1997.

[VC64] V. Vapnik and A. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.

[VC71] V. Vapnik and A. Chervonenkis. On the uniform covergence of relative frequencies of events to their probabilities. *Theory Probab Appl.*, 16:264–280, 1971.

[VL63] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.

[WW97] M.O. Stitson, A. Gammerman, V. Vapnik, C. Watkins and J. Weston. Support vector anova decomposition. *Technical Report Royal Holloway College*, CSD-TR-97-22, 1997.